

UNITED STATES PATENT APPLICATION

of

Jacob M. McGuire

for a

HIGHLY, SCALABLE LEAST CONNECTIONS LOAD BALANCING

HIGHLY, SCALABLE LEAST CONNECTIONS LOAD BALANCING

CROSS-REFERENCE TO RELATED APPLICATIONS

The present application claims priority from U.S. Provisional Patent Application 5 Serial No. 60/236,555, which was filed on September 29, 2000, by Jacob M. McGuire for a HIGHLY, SCALABLE LEAST CONNECTIONS LOAD BALANCING and is hereby incorporated by reference.

BACKGROUND OF THE INVENTION

Field of the Invention

10 The present invention relates to load balancing traffic among a plurality of servers, and in particular, to a least connections load balancing method.

Background Information

Computer networks typically use file servers which frequently operate under a client-server paradigm. Under this model, multiple clients can make input/output (I/O) 15 requests which are directed to a particular resource on the network. A server on the network receives and carries out the I/O requests. When the server receives multiple I/O requests, the server may choose to service them one at a time. I/O requests which are not being processed typically wait until the server is ready to receive more requests. As a 20 result, the server can become a bottleneck in the network.

Typically, it is desirable to distribute various client requests among the plurality 20 of servers. In these instances, it requires collaboration as to how the various client requests are to be distributed among those various servers. This may be performed through

load balancing. Server load balancing allows for a group of real servers (a server farm) to be represented as a single virtual server entity wherein the traffic is balanced among the plurality of servers. One method to obtain server load balancing is to use a round-robin method. With this method, a new connection between a client and a real server is 5 performed by choosing the real servers in a circular manner wherein a connection is made if the chosen server has capacity to handle the connection. However, the round-robin method does not insure that the various real servers are indeed effectively load balanced.

Another method for server load balancing is a least connections method in which a new connection is assigned to the real server with the least number of currently active 10 connections. Compared with the round-robin method, the least connections method provides for a more accurate load balancing of the servers; however, it is rather complex and consumes a fair amount of processing time of the device that is performing the load balancing. For instance, the method sends a new connection to a server which has the lowest metric, wherein the metric is defined as the number of connections on the server divided 15 by the weight (or capacity) of the server. This metric is kept as a quotient/remainder pair. To keep track of the metric and the remainder, integer division is typically performed on all servers every time a connection is added or removed.

On a different note, it is desired to flexibly increase the number of real servers as demand increases for resources at the server farm. However, one aspect of this problem 20 is that as greater numbers of real servers are added, the load balancing process slows down, which may require replacing the load balancing device with one that processes at a faster speed. For instance, consider a method in which the load balancing device sequentially tests a list of servers for least connections. One reason for the slowdown may be that as the list of potential servers for connections increase, more time is needed for the 25 device to find the server with the lowest number of active connections.

SUMMARY OF THE INVENTION

A load balancing device according to an embodiment of the invention uses a predictor that comprises a plurality of Least Connections Control Blocks (LCCBs) configured to keep track of the real servers with active connections. To speed up the search for

a real server with the least number of active connections, an LCCB is kept for each server metric. A metric is defined as the number of connections on a server divided by the weight (or capacity) of the server. This metric is kept as a quotient/remainder pair. The predictor sends out the real server address with the lowest metric from the LCCB with the least connections whenever a new connection is required by the load balancing device.
5 According to the embodiment, a list of metric/pointer pairs is kept by the associated LCCB, one corresponding to each real server, and when a connection is requested, the real server with the lowest metric is used for the connection which is pointed to by the pointer of the metric/pointer pair.

10 Each of the LCCBs is associated with a list of server metric/pointers that points to servers having a similar number of connections. Because the LCCBs are kept in sorted order (in terms of the number of connections in a server), finding the least loaded server is simply a matter of selecting the LCCB with the lowest connections and taking the first real server with the lowest metric (which is pointed to by the LCCB). The metric of the server is revised and the server metric/pointer is transferred to another LCCB which corresponds to the server's new number of connections. In particular, when a connection is added to the server, the remainder (which is also called a "slope") is incremented, and when a connection is removed, the slope is decremented. When the slope equals the weight, or goes below zero, the metric is updated and the slope reset.
15

20 As the number of connections of a real server changes, the server metric is removed from its current LCCB. Because the predictor creates/destroys connections sequentially, the request that a connection be created or destroyed occurs at most one at a time. Thus, if the connection of the real server changes, at most, the server metric/pointer is moved to one of the immediate LCCBs. For example, if a server metric/pointer is on the list of the LCCB having two connections and another connection is added to it, the metric is revised and that server metric/pointer is moved to an adjacent LCCB which is associated with three connections. The server metric/pointer is transferred to the new LCCB via a "double linked" connector, wherein the new LCCB subsequently updates and sorts its server metric/pointer list. The new LCCB then points to the server with the lowest metric. Conversely, if a server in the LCCB with two connections loses a connection,
25
30

the metric is revised and the metric/pointer is moved to the adjacent LCCB which is associated with one connection. The LCCB transfers the server metric/pointer to the new LCCB via the double linked connector. The new LCCB then updates and sorts its server metric/pointer list and points to the server with the lowest metric.

5

BRIEF DESCRIPTION OF THE DRAWINGS

10

The invention description below refers to the accompanying drawings, of which:

Fig. 1 illustrates a network in which the present invention may be implemented;

Fig. 2 illustrates a virtual server including a server farm and a load balancing device;

Fig. 3 is a schematic diagram of the load balancing device in which an embodiment of the invention may be implemented;

Fig. 4 is a schematic diagram of a processing unit of the load balancing device constructed in accordance with an embodiment of the invention;

15

Fig. 5a is a predictor constructed in accordance with an embodiment of the invention; and

Fig. 5b is the predictor of Fig. 5a in which a redundant Least Connections Control Block (LCCB) has been removed in accordance with an embodiment of the invention.

20

DETAILED DESCRIPTION OF AN ILLUSTRATIVE EMBODIMENT

25

Fig. 1 shows an illustrative network 100 including a network cloud 110 which may be a combination of transmission lines, backbones, switches, routers and repeaters, dependent on the extent of the network, and allows various devices connected to the network to communicate with each other. These various devices may include simple computers such as personal computers (PCs) and workstations 112, (herein referred to as clients) that transmit or receive data, or they may be sophisticated servers 114 that store various resources. The network 100 may also include virtual servers 200, which operate in accordance with an embodiment of the present invention, to exchange data with clients

112. The network cloud 110, for instance, may be the Internet that interconnects large numbers of these simple computers, servers and virtual servers to facilitate the exchange of information. Information, in one example, flows in packets having attached headers that include, among others, source and destination addresses that inform the devices encountered in the network as to how the packets are to be forwarded so that they may reach their destination.

5 **Fig. 2** is an exemplary virtual server 200 that includes a server farm 201 and a load balancing device 300 that provides for server load balancing and allows for clients to be represented by a group of network servers. From the clients' perspective, the virtual 10 server 200 is a single server entity. One purpose of the virtual server is to facilitate scalability and to expedite resource availability for the clients. Moreover, the addition of new servers and/or the removal/failure of existing servers can occur at any time without affecting the resource availability provided by the virtual server. Server load balancing allows for traffic to be balanced among the servers and in certain instances, limit traffic to 15 certain clients or servers. The servers that comprise the server farm 201 are herein referred to as "real servers" 202.

20 According to one embodiment, the clients 112 are configured to communicate with the virtual server 200 using an IP address that represents the virtual server (herein-after referred to as "virtual IP address."). When a client 112 initiates a connection with the virtual server 200, this connection is received by a load balancing device 300 which, in turn, chooses a real server 202 from the server farm 201 using the novel load balancer (to be described herein). Although the load balancing device 300 is shown as being separate from the server farm 201, in another embodiment, it may be integrated into one of the servers 202 in the server farm 201.

25 **Fig. 3** is a diagram of an illustrative load balancing device 300, such as a switch, that directs traffic between the clients 112 and the real servers 202. The load balancing device comprises a processing unit 400, a forwarding engine 304 and a plurality of input/output (I/O) ports 340. The processing unit 400 may be a single processor or a plurality of processors that manage the load balancing device and execute, among other 30 things, routing protocols. The processing unit 400 also executes a load balancing algo-

rithm according to one embodiment of the invention. The forwarding engine 304 is a switch component that performs packet switching of packets received from the I/O ports 340, some of which are connected to the network cloud 110 and others of which are connected to the server farm 201. The load balancing device 300, in one embodiment, performs packet redirection (in another embodiment, it may be a packet translation) to and from the real servers 202 in the server farm 201 on traffic from clients requested through the virtual IP address. The load balancing device 300 uses a unique real server address to address a specific real server.

Communication between a client and a server farm is established by a “connection.” A connection is a series of Internet Protocol (IP) packets associated with the same pair of IP addresses and the same Transport Control Protocol/User Datagram Protocol (TCP/UDP) ports. The information that describes packets belonging to a unique client/server connection and allows for a packet to be sent to and from a real server includes:

- 15 Source IP address;
- Destination IP address;
- Protocol;
- Source TCP/UDP port;
- Destination TCP/UDP port; and
- 20 IP address to forward the packet to.

The following example describes how a load-balanced TCP connection between a client and a server is processed by the load balancing device, which in this instance is a switch. The TCP protocol is connection-oriented and has known protocol messages for activation and deactivation of TCP sessions. Thus, under one method, a state machine is used to correlate TCP packets such as SYN, SYN/ACK, FIN and RST for realizing a connection is to be added/removed or for determining the number of connections per server. To expedite the load balancing process, an Access List Control 320 may be used to filter the TCP packets that indicate that a load balancing process is to take place.

During initiation of a connection, the client sends a TCP SYN packet using the virtual IP address. When the switch 300 receives this packet, it is passed through the Access List Control 320, wherein logic encoded as an Application Specific Integrated Circuit 1 (ASIC1) 322 compares the packet's header with an Access Control List (ACL) 324. The ACL 324 may be stored in a Random Access Memory (RAM) or a Content Addressable memory (CAM). The ACL 324 looks at the packet header to determine if the packet should be in the embodiment where the load balancer is constructed using ASIC technology and implemented in the forwarding engine 304, the SYN packet is sent to the load balancer 310 to be processed. In an alternative embodiment, where the processing unit 400 performs the load balancing, the SYN packet is marked with a software index that indicates that it is to be load balanced and is sent to the processing unit. Whether the load balancer is in the forwarding engine 304 or is in the processing unit 400 or is a separate component in the switch, its operation is similar. Thus, the embodiment in which the load balancer is implemented in a processing unit will be described herein. As the packet is redirected to the processing unit 400, the ASIC2 332 of the forwarding unit 330 creates a new flow instance in its forwarding table 334. A flow instance is an entry in the forwarding table that indicates where the packet should be forwarded to.

Referring to **Fig. 4**, when the processing unit 400 receives the packet, the processor 402 sees that the index corresponds to load balancing, and processes the packet using a load balancer 410, which may be implemented in hardware or software. The load balancer 410 maps this index to a virtual server instance, decodes the packet and searches a connection database 412 to find a connection. If it fails to find a connection, the load balancer 410 chooses a real server (i.e., the server address) using a predictor 500 (which may be implemented in hardware or software) and creates a connection in the connection database 412. The predictor 500 will be further described herein. The chosen real server address is conveyed to the forwarding engine 304 as the SYN packet is placed on the bus 302. The forwarding engine 304 then "caches" the server address in the forwarding table 334 as the SYN packet is forwarded to the chosen real server.

The chosen server 202 responds with a TCP SYN/ACK packet. This packet is received by the switch 300 and is passed to the Access List Control 320. The ASIC1 322

redirects the packet to the processing unit 400 while marking it with a software index that indicates that it pertains to load balancing. Meanwhile, ASIC2 332 creates a flow instance in its forwarding table 334. Note that the forwarding unit 330 keeps two separate caches; one for the inbound packets and one for the outbound packets. Of course, this
5 may be modified according to a desired result. The processor 402 in the processing unit 400 processes the packet using the load balancer 410 which maps this index to a virtual server instance, decodes the packet and finds the connection in the connection database 412. The client address is thus found and the packet is placed on the bus 302. Logic currently encoded on ASIC2 332, meanwhile, learns the client address for this flow and
10 caches this in the forwarding table 334 as the SYN/ACK packet is forwarded to the client.

The client then sends a TCP ACK packet. When the switch 300 receives this packet, the ASIC1 322 does not find a match in the ACL 324 and permits the packet to be forwarded as opposed to redirecting it to the processing unit 400. Because the forwarding unit 330 has a flow entry in its forwarding table 334 for this flow, the forwarding engine 304 sends the packet directly to the associated real server 202.
15

TCP Data packets are now exchanged between the client and the server during this TCP session. Because the flags of interest i.e., SYN, FIN, and RST are not set on any of these packets, and the forwarding unit 330 has the flow entries in its forwarding
20 table 334, the forwarding engine 304 sends the packets directly to the server or client without consulting the processing unit 400 (i.e., the load balancer).

Towards the end of the session, the client sends a TCP RST packet. When the switch receives this packet, the ASIC1 322 consults the ACL 324 to determine if it needs to be redirected. If so, the ASIC1 322 redirects the RST packet to the processing unit 400
25 while marking it with a software index that indicates that it pertains to load balancing. The processing unit 400 receives this packet and the processor 402 sees that the index corresponds to load balancing and maps this index to a virtual server instance, decodes the packet and finds the connection in the connection database 412. The connection is then removed from the connection database 412 and the memory space consumed by the
30 connection is returned to the free connection pool.

The present invention also applies to a situation wherein the client sends a UDP packet. Because UDP is not a connection-oriented protocol, UDP protocol messages cannot be “sniffed” (without knowing details of the upper-layer protocol) to detect the beginning or the end of a UDP message exchange. In this instance, the detection of the 5 UDP connection may be based on a configurable idle timer.

When the client issues a UDP packet using the virtual IP address and the switch 300 receives this packet, ASIC1 322 permits the packet to be forwarded since no matching entry exists in the ACL 324. Because the forwarding unit 330 does not have a corresponding flow in its forwarding table 334, it causes the ASIC2 332 to create a new flow 10 and to forward the packet to the processing unit 400 where a lookup operation is performed to locate the route. The processing unit 400 receives this packet and because there is no software index that indicates load balancing, the processor consults a Common Packet Filter (CPF) 404. The CPF 404 forwards this packet to the load balancer 410 because it has a matching and best-fit packet filter. The load balancer 410 maps this index 15 to a virtual server instance, decodes the packet, searches the connection database 412, and fails to find a connection. The load balancer then uses a predictor 500 to choose a real server, creates a connection, inserts it into the connection database 412, and starts the connection's idle timer. The real server IP address is found and this is used forward the packet on the bus 302. The ASIC2 332 learns the server address and caches this in the 20 forwarding table 334 as the packet is forwarded to the chosen real server 201.

The chosen server 202 responds with a UDP packet. When the switch 300 receives this packet, the ASIC1's 322 decision is to simply permit the packet to be forwarded because there is no matching entry in the ACL 324 that indicates that it is to be load balanced. Because the forwarding unit 330 does not have a corresponding flow in 25 its forwarding table 334, it causes the ASIC2 332 to create a new flow and forwards the packet to the processing unit 400 to lookup the route. The processing unit 400 receives this packet and since there is no software index that indicates that it is to be load balanced, the processor 402 causes the packet to be handled by the CPF 404. The CPF 404 forwards this packet to load balancer 410 since it has a matching and best-fit packet filter. 30 The load balancer 410 maps this index to a virtual server instance, decodes the packet,

finds the connection in the connection database 412, and ascertains the client's IP address. This IP address is used to forward the packet on the bus 302. Meanwhile, the ASIC2 332 learns the client's IP address while the packet is forwarded to the client.

5 UDP Data packets are exchanged between the client and the server using the same set of ports. Since ASIC1 322 simple permits the packets and the forwarding unit 330 has the flow entries in its forwarding table 334, the forwarding engine 304 sends the packets directly to the server or client. Eventually, the UDP connection will become idle for a period of time exceeding the idle timer. At this point, the load balancer 410 will timeout the connection.

10 **Fig. 5a** shows an exemplary predictor 500 constructed in accordance with an embodiment of the invention. The predictor 500 may be constructed using hardware (such as ASIC) or software. The predictor comprises a plurality of Least Connections Control Blocks (LCCBs) 502 that keeps track of real servers with active connections. To speed up the search for the real server with the least number of active connections, an LCCB is 15 kept for each server metric. A metric is defined as the number of connections on a server divided by its weight (or capacity) of the server. This metric is kept as a quotient/remainder pair. The predictor 500 sends out via a selector 508 the real server address with the lowest metric from the LCCB with the least connections whenever a new connection is required by the load balancer 410. According to the embodiment, a list of 20 metric/pointer pairs 504 is kept in a memory of the associated LCCB, one corresponding to each real server 202, and when a connection is requested, the real server with the lowest metric/pointer pair 504 is used for the connection which is pointed to by the pointer of the metric/pointer pair.

25 Each of the LCCBs is associated with a list of server metric/pointers that points to servers having similar number of connections. Because the LCCBs are kept in a sorted order (in terms of the number of connections in a server), finding the least loaded server 202 is simply a matter of the selector 508 selecting the LCCB with the lowest connection and taking the first real server with the lowest metric (represented by the metric/pointer pair 504) which is pointed to by a pointer 506 of that LCCB. As the real server address is 30 supplied to the load balancer, the metric of the server is revised and the server met-

ric/pointer 504 is transferred to another LCCB (while that metric/pointer pair is removed from the LCCB’s metric/pointer pairs list). The LCCB then points to the next server metric/pointer 504 on the list. The server metric/pointer is transferred to another LCCB which corresponds to the server’s number of connections. In particular, when a connection is added to the server, the remainder (which is also called a “slope”) is incremented, and when a connection is removed, the slope is decremented. If the slope becomes equal to the weight, or goes below zero, the metric is updated and the slope reset.

As the number of connections of a real server changes, its corresponding metric/pointer is removed from its current LCCB. Because the predictor creates/destroys a connection sequentially, the request that a connection be created or destroyed occurs at most one at a time. Thus, if the connection of the real server changes, at most, the server metric/pointer is moved to one of the immediate LCCBs. For example, if a server metric/pointer is on the list of the LCCB with two connections and another connection is added to the server, the server metric/pointer is revised and is moved to the adjacent LCCB associated with three connections. The server metric/pointer is transferred to the new LCCB via the double linked connector, wherein the new LCCB then updates and sorts its server metric/pointer list. The new LCCB then points to the server with the lowest metric. Conversely, if a server with a two connection loses a connection, its corresponding metric/pointer pair is revised and is moved to the adjacent LCCB which is associated with one connection. The LCCB transfers the server metric/pointer to the new LCCB via the double linked connection. The new LCCB then updates and sorts its server metric/pointer list and points to the server with the lowest metric.

To further expedite the process and to conserve resources, it is desired that for LCCBs that do not have an active server metric/pair to be removed altogether to conserve memory allocation. In the case of the load balancing device that is implemented in hardware, the redundant LCCBs can be deactivated (i.e., powered down). Take the example where the server with one connection has lost that connection. Its corresponding metric/pointer is transferred to the LCCB with no connections. In this instance, the LCCB with one connection no longer has any server metric/pointer to keep track of. In such cases, it is desired to remove/shutdown the LCCB to conserve memory allocation and

software processing time, the result which is shown as **Fig. 5b**. Referring to **Fig. 5b**, in the event that the LCCB with one connection has a connection added to one of the servers pointed to by a metric/pointer on its list, that server now has two connections. However, the LCCB with two connections does not exist. In this instance, a new LCCB with two 5 connections is created and the metric/pointer of the real server is added to the list of that control block. In the case where the load balancing device is implemented in hardware, the LCCB can be powered up to receive the metric/pointer pair.

As a performance enhancement we attach the current "best real server" and its metric and only search for a new real server when the metric of the current best real 10 server increases or the metric of another real server is incremented below the current metric of the real server. Note that new servers may be added or a failed server may be removed without affecting the performance of the predictor and thus the load balancer. This is because finding the least loaded server is a simple matter of looking at the first LCCB with the least connections and finding the server pointed to by the LCCB.

15 A method and apparatus of providing a highly scalable least connections load balancing has been described. It will however be apparent that other variations and modifications may be made to the described embodiment, with the attainment of some or all of their advantages. Therefore, it is the object of the appended claims to cover all such variations and modifications that come within the true spirit and scope of the invention.

20 What is claimed is: